Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making

Year: 2020 | Citations: 770 | Authors: Yunfeng Zhang, Q. V. Liao, Rachel K. E. Bellamy

Abstract

Today, AI is being increasingly used to help human experts make decisions in high-stakes scenarios. In these scenarios, full automation is often undesirable, not only due to the significance of the outcome, but also because human experts can draw on their domain knowledge complementary to the model's to ensure task success. We refer to these scenarios as AI-assisted decision making, where the individual strengths of the human and the AI come together to optimize the joint decision outcome. A key to their success is to appropriately calibrate human trust in the AI on a case-by-case basis; knowing when to trust or distrust the AI allows the human expert to appropriately apply their knowledge, improving decision outcomes in cases where the model is likely to perform poorly. This research conducts a case study of AI-assisted decision making in which humans and AI have comparable performance alone, and explores whether features that reveal case-specific model information can calibrate trust and improve the joint performance of the human and AI. Specifically, we study the effect of showing confidence score and local explanation for a particular prediction. Through two human experiments, we show that confidence score can help calibrate people's trust in an AI model, but trust calibration alone is not sufficient to improve AI-assisted decision making, which may also depend on whether the human can bring in enough unique knowledge to complement the AI's errors. We also highlight the problems in using local explanation for AI-assisted decision making scenarios and invite the research community to explore new approaches to explainability for calibrating human trust in AI.